Systematic benchmarking of detection tools for CpG methylation from Nanopore sequencing

Zaka Wing-Sze Yuen[1,2], Cameron Jack[1], Eduardo Eyras[1,2]

[1] The John Curtin School of Medical Research, Australian National University, Acton ACT 2601, Australia
[2] EMBL Australia Partner Laboratory Network at the Australian National University, Acton ACT 2601, Australia

This study benchmarked five different detection tools for modified base from Nanopore sequencing – **Nanopolish[1], Megalodon, DeepSignal[2], Guppy and Tombo[3]**, for their accuracy to predict DNA CpG methylation state from Nanopore sequencing.

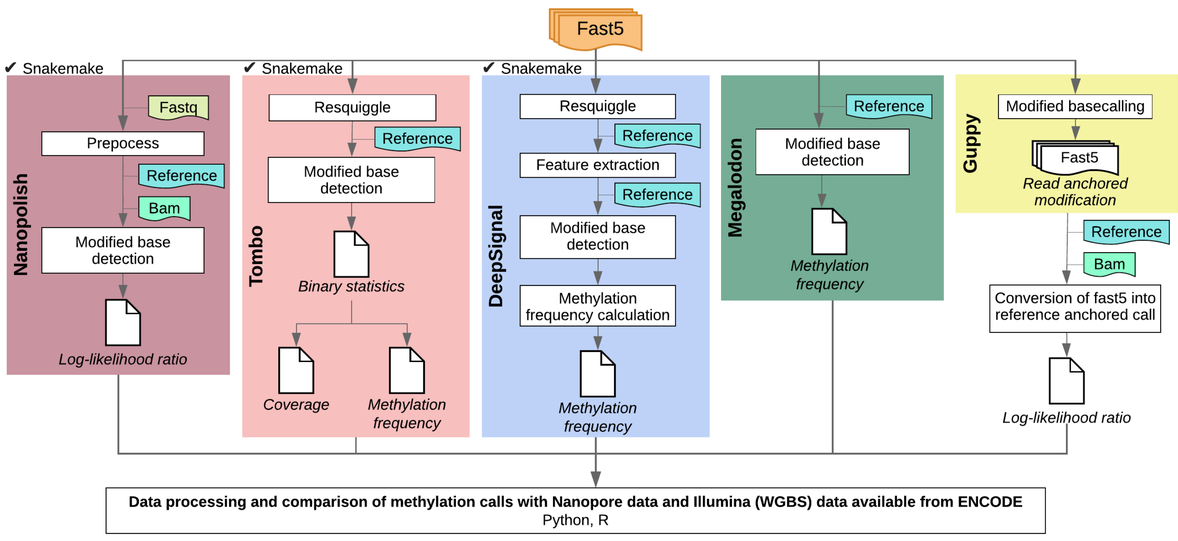
INTRODUCTION

Existing methods for modified base detection using require either immunoprecipitation of DNA or chemical conversion like for example bisulfite treatment, lacking enough resolution and sensitivity. Nanopore sequencing can access native DNA at single-molecule resolution and detect base modifications from the Nanopore signal patterns. Despite the various detection tools available for modified bases from Nanopore sequencing, there has not been yet a systematic accuracy benchmarking to determine the strengths and limitations of these approaches.

METHODS

We performed a PCR-free targeted nanopore sequencing experiment using the Cas9 enrichment approach to preserve modifications and to capture 10 different regions in a human cell line NA12878, and the DNA library was sequenced on a single MinION flowcell.

Reproducible pipelines for CpG methylation detection



Downstream processing of Guppy's output: https://github.com/kpalin/gcf52ref

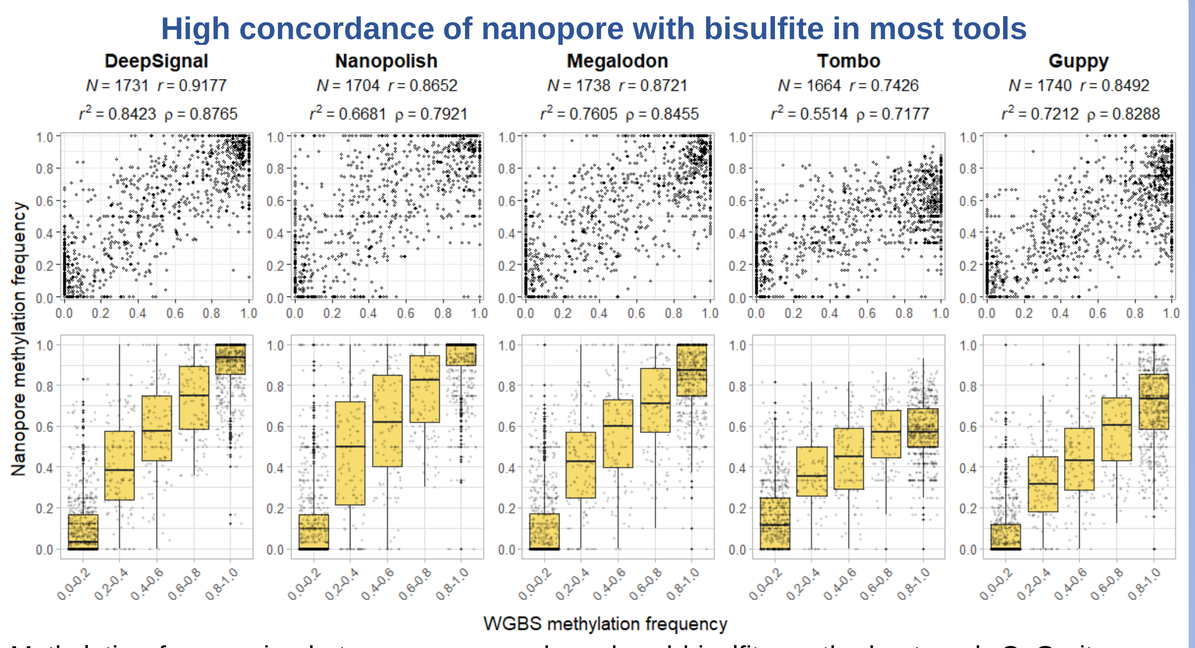
Since there are multiple steps in modified base detection for Nanopolish, DeepSignal and Tombo, we have developed a set of Snakemake pipelines to make these analyses reproducible. Post-processing for the methylation calling results was then performed to generate a methylation frequency for all mapped CpGs on both strands.

To obtain high confidence methylation calls from published whole-genome bisulfite sequencing (WGBS) data for subsequent validation, the CpG sites in the enriched regions with zero coverage from two WGBS replicates or difference in methylation frequency between two replicates above 0.9 quantile and below 0.1 quantile were removed. The methylation frequency of each CpG from all tools was compared to that from the Illumina BS-based data for the same individual.

We also applied the same pipeline to the high-coverage enrichment data from a recently published study[4] performing nanopore Cas9-targeted sequencing. In addition to validation with WGBS data, we used the *E.coli* methylation control datasets[1] as the ground truth to establish accuracy of each tool.

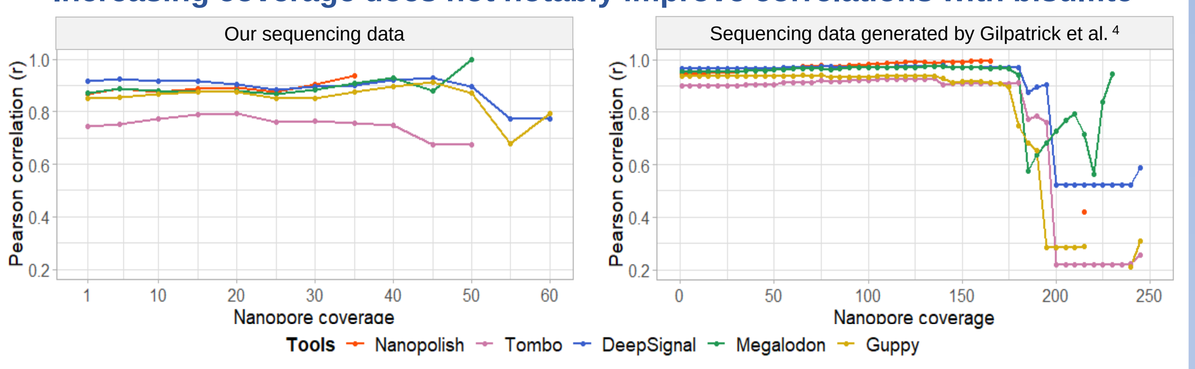
RESULTS

High concordance of nanopore with bisulfite in most tools



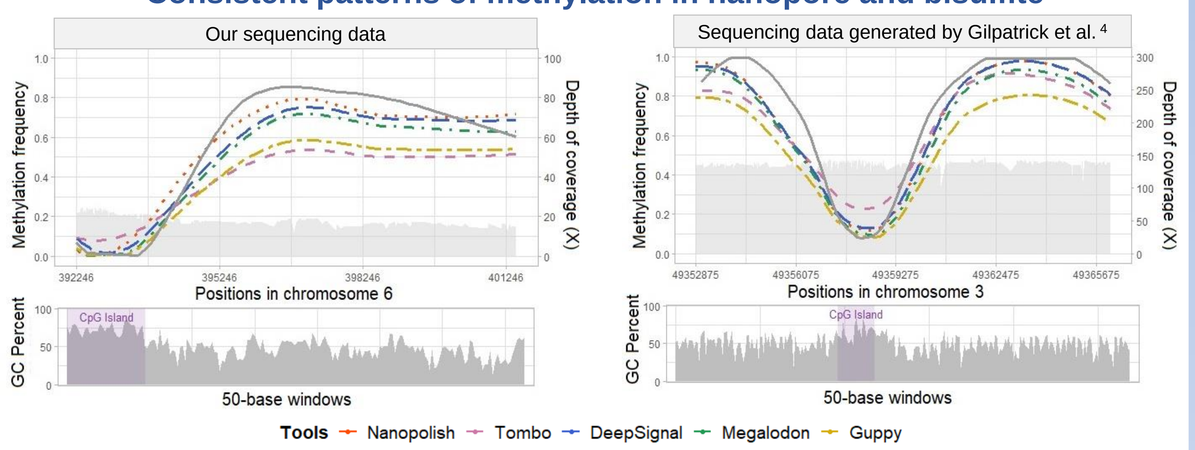
Methylation frequencies between nanopore-based and bisulfite methods at each CpG site were compared using the Pearson and Spearman correlations. DeepSignal showed the highest correlation (r = 0.92) among the five tools followed by Megalodon (r = 0.87), Nanopolish (r = 0.87) , Guppy (r = 0.84) and Tombo (r = 0.74). To compare the spread of methylation frequency distribution, we further catergorised CpG sites from bisulfite data into five bins with increasing frequencies. DeepSignal had a more consistent uphill pattern and lower variability than the other four tools across the bins.

Increasing coverage does not notably improve correlations with bisulfite



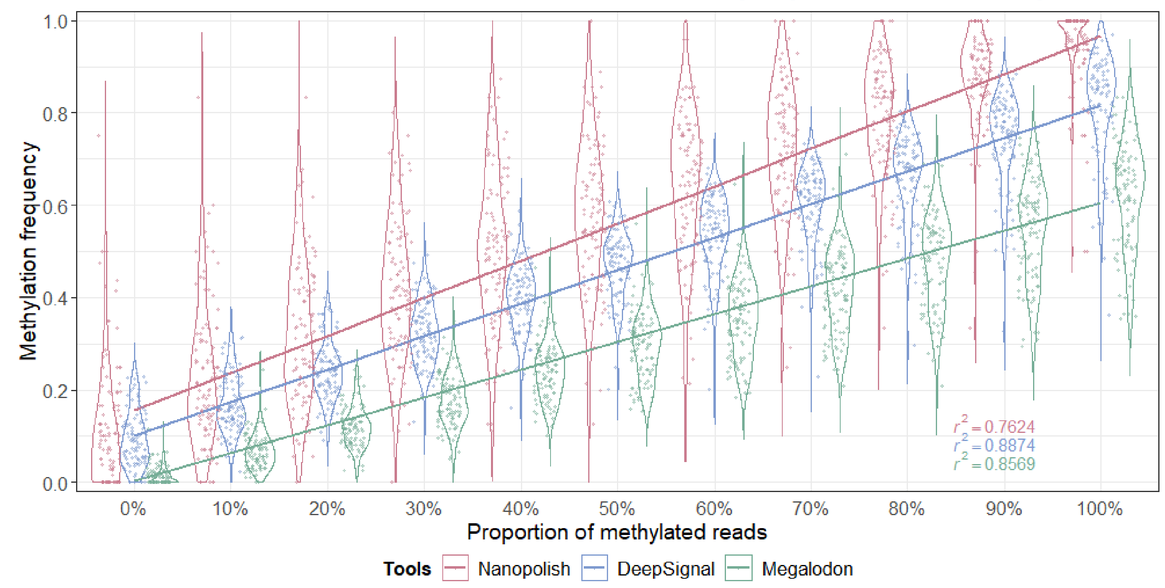
All tools suffered from systematic biases that failed to predict methylation state at high coverage.

Consistent patterns of methylation in nanopore and bisulfite



Methylation patterns were concordant between nanopore-based and bisulfite methods, even at low coverage.

Distribution for control datasets with different proportions of methylation



Various mixes with different proportions of methylated reads were created using the methylation control datasets[1]. So far only the results of Nanopolish, DeepSignal and Megalodon have been generated, and more results will be released soon. In 0% methylated (negative control), Megalodon was more accurate and had fewer false positives; In 100% methylated (positive control), Nanopolish was more accurate and had fewer false negatives. Using the linear regression model, the R squared of DeepSignal was the highest, which explained 88.7% of variance.

REFERENCES

1. Simpson et al. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature methods*, 14(4), 407.
2. Ni et al. (2019). DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22), 4586-4595.
3. Stoiber et al. (2016). De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. *BioRxiv*, 094672.
4. Gilpatrick et al. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature biotechnology*, 38(4), 433-438.

ACKNOWLEDGEMENTS

We would like to thank Oxford Nanopore Technologies for travel bursaries; Timothy Gilpatrick, Jared Simpson and the ENCODE Consortium for generating and making the GM12878 enrichment data, the E.coli methylation data and the GM12878 WGBS data publicly available respectively.

Australian National University

EMBL Australia

zaka.yuen@anu.edu.au

zakayuen21